Research Report

A Unified Approach to IRT Scale Linking and Scale Transformations

Matthias von Davier

Alina A. von Davier

Research & Development

March 2004
RR-04-09

A Unified Approach to IRT Scale Linking and Scale Transformations

Matthias von Davier and Alina A. von Davier

ETS, Princeton, NJ

March 2004

Research Reports provide preliminary and limited dissemination of ETS research prior to publication. They are available without charge from:

Research Publications Office
Mail Stop 7-R
ETS
Princeton, NJ 08541

Abstract

This paper examines item response theory (IRT) scale transformations and IRT scale linking methods used in the Non-Equivalent Groups with Anchor Test (NEAT) design to equate two tests, X and Y. It proposes a unifying approach to the commonly used IRT linking methods: mean-mean, mean-var linking, concurrent calibration, Stocking and Lord and Haebara characteristic curves approaches, and fixed-item parameters scale linkage. The main idea is to view any linking procedure as a restriction on the item parameter space. Then a rewriting of the log-likelihood function together with an appropriately implemented maximization procedure of the log-likelihood function under linear (or nonlinear restrictions) will accomplish the linking. The proposed method is general enough to cover both the dichotomous item response models (the one parameter logistic (1PL) model, 2PL, and 3PL) and the polytomous unidimensional IRT models like the generalized partial credit model.

Key words: Item response models, scale transformation, test linking, Non-Equivalent Groups with Anchor Test Design, nonlinear restrictions, maximization procedures, Lagrange multipliers

Acknowledgements

The authors thank Shelby Haberman, Hariharan Swaminathan, Henry Braun, and Paul Holland for fruitful discussions and insight, and Wendy Yen, Neil Dorans, and Dan Eignor for their feedback and suggestions on the previous draft of this paper.

1. Introduction

The need for equating arises when two or more tests on the same construct or subject area can yield different scores for the same examinee. The goal of test equating is to allow the scores on different forms of the same tests to be used and interpreted interchangeably. Item response theory (IRT, Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980; Thissen & Wainer, 2001; and many others) has provided new ways to approach test equating. If IRT is used in the equating process, it is usually also necessary to use some sort of linking procedure to place the IRT parameter estimates on a common scale.

In this paper we focus on the IRT linking procedures used for data collection designs that involve "common items." The data collection designs that use this method are called here "Non-Equivalent groups with Anchor Test (NEAT)" designs and can have both internal and external anchor tests (see, for example, von Davier, Holland, & Thayer, 2004; Kolen & Brennan, 1995).

In the NEAT design there are two populations, P and Q, of test-takers and a sample of examinees from each. The sample from P takes test X, the sample from Q takes test Y, and both samples take a set of common items, the anchor test V. This design is often used when only one test form can be administered at one test administration because of test security or other practical concerns. The two populations may not be "equivalent" (i.e., the two samples are not from a common population).

The two tests X and Y and the anchor V are, in general, not *parallel*. The anchor test V is usually shorter and less reliable than either X or Y. Angoff (1984/1971) gives advice on designing anchor tests. For a comparison of a variety of methods for treating the NEAT design, see Kolen and Brennan (1995), Marco, Petersen, and Stewart (1983), and Petersen, Marco, and Stewart (1982).

In this paper we examine the IRT scale transformation and IRT linking methods used in the NEAT design to link X and Y. More exactly, we propose a unified approach to the IRT linking methods: mean-sigma and mean-mean, concurrent calibration, fixed parameters calibration, the Stocking and Lord characteristic curves approach, and the Haebara characteristic curves approach (see Kolen & Brennan, 1995, chapter 6, for a

detailed description of these methods). Moreover, we believe that our view of IRT linking can be extended to cover other flavors of IRT scaling/linking procedures.

In our approach, the parameter space is described by all the parameters of the IRT model fitted to the data from both populations in a marginal maximum likelihood (MML) framework. Under the usual assumptions for the NEAT design, which are described later, the joint log-likelihood function for this model on the data from both populations can be expressed as the sum of two log-likelihood functions corresponding to each of the two groups of data and parameters.

The main idea in our approach is to view *any linking method* as *a restriction function* on the joint parameter space of the instruments to be equated.

Once this is understood, rewriting the joint log-likelihood function by including a term for each restriction and an appropriately implemented maximization procedure will accomplish the linking. The maximization is carried out using a vector of Lagrange multipliers (see, for example, Aitchison & Silvey, 1958; von Davier, 2003; Glas, 1999).

We will show that the new approach is general enough to cover the usual item response models (IRMs: the one parameter logistic (1PL), 2PL, and 3PL models) as well as polytomous unidimensional IRT models like the generalized partial credit model (GPCM).

Some of the advantages of this new perspective on IRT linking are

- providing a common framework for all IRT scale linking methods yields a better understanding of the differences between the approaches, which opens paths to more flexible methods of IRT linking;

- viewing the IRT linking as a restriction function allows us to control the strength of the restriction. For example, the concurrent calibration is the most restrictive IRT linking method, as it assumes the equality of all parameters in the anchor test. When such a strong restriction is not appropriate, the proposed method provides alternatives;

- providing a *family* of linking functions that ranges from the most restrictive one, the concurrent calibration, to separate calibration (without additional restrictions, i.e., to no linking at all);

- possibly allowing the implementation of a hierarchical structure of the item parameters in the anchor in a way that might be useful for vertical linking. Such a hierarchical structure was proposed by Patz, Yao, Chia, Lewis, and Hoskens, 2003; their estimation method was Markov chain Monte Carlo (MCMC). Later in this paper we propose an analytical approach. We also think that the approach considered here can be extended to multidimensional IRT models. However, this is an interesting future research topic; and

- allowing the development of statistical tests (such as Lagrange multiplier tests) for checking the appropriateness of different IRT linking methods (following similar principles as in Glas, 1999).

In this paper, we focus on the description of the theoretical framework and derivations of a new approach to IRT linking. The rest of the paper is structured as follows: First we introduce our notation; we briefly describe the well known IRT linking methods in the next section. Then we investigate the joint log-likelihood function and the restriction function more formally and for several IRT linking methods. Finally we discuss the advantages of this perspective on the IRT linking.

2. The NEAT Design and IRT Linking

2.1. The NEAT Design

The data structure for the NEAT design is illustrated in Table 1 (see also von Davier et al., 2004).

Table 1
Non-Equivalent Groups With Anchor Test (NEAT) Design

Population	Sample	X	Y	V
P	1	$\surd$		$\surd$
Q	2		$\surd$	$\surd$

Note that Table 1 describes the data collection procedure and does not refer to the test scores, as might be the case in observed-score equating. We will denote the matrices of

observed item responses to the tests X, V, and Y by $\boldsymbol{X}$, $\boldsymbol{V}$, and $\boldsymbol{Y}$. The subscripts P and Q will denote the populations.

The analysis of the NEAT design usually makes the following assumptions (see also von Davier et al., 2004):

Assumption 1. *There are two populations of examinees P and Q, each of which can take one of the tests and the anchor.*

Assumption 2. *The two samples are independently and randomly drawn from P and Q, respectively.*

Table 1 shows that in the NEAT design $\boldsymbol{X}$ is not observed in population Q, and $\boldsymbol{Y}$ is not observed in population P. To overcome this feature, all linking methods developed for the NEAT design (both observed-score and IRT methods) must make additional assumptions of a type that does not arise in the other linking designs.

Assumption 3. *The tests to be equated, X and Y, and the anchor V are all unidimensional (i.e., all items measure the same unidimensional construct), carefully constructed tests, in which the local independence assumption holds (Hambleton et al., 1991).*

These three assumptions are sufficient for our exposition. We will not make any distributional assumptions.

2.2. Unidimensional IRT Models

IRT models rely on the assumptions of monotonicity, undimensionality, and local independence (Hambleton et al., 1991). These models express the probability of a response x_{ni} of a given person, n ($n = 1, \ldots N$), to a given item, i ($i = 1 \ldots, I$), as a function of the person's ability (latent), θ_n, and a possibly vector valued item parameter, $\boldsymbol{\beta}_i$, that is,

$$P_{ni} = P(X = x_{ni}) = f(x_{ni}, \theta_n, \boldsymbol{\beta}_i)$$

In the case of the well known 3PL model (Lord & Novick, 1968), the item parameter is three-dimensional and consists of the slope, the difficulty, and the guessing parameter, that

is, $\boldsymbol{\beta}_i^t = (a_i, b_i, c_i)$.

The 3PL model, which serves as the standard example of an IRM in this paper, is given by

$$P\left(x_i = 1 \mid \theta, a_i, b_i, c_i\right) = c_i + (1 - c_i)\text{logit}^{-1}\left[a_i\left(\theta - b_i\right)\right],$$

with $\text{logit}^{-1}(x) = \exp(x)/(1 + \exp(x))$.

However, most results presented here do not depend on the specific choice of the IRM and apply to models for both dichotomous and polytomous data.

2.3. IRT Linking

When conducting scale linking in the NEAT design, the parameters of the IRM from different test forms need to be on the same IRT scale. If the calibration was carried out separately on the two samples from the two different populations P and Q, then parameter estimates for the anchor test will be available for examinees in the two groups. These separate parameter estimates of the anchor in the two groups serve as the basis for the scale transformation (mean-mean, mean-sigma methods, or characteristic curves methods, such as Stocking and Lord or Haebara methods).

As an alternative, the item parameters from X, V (in both populations), and Y can be estimated jointly, coding the items that an examinee did not take as "not administered" or "not reached," since these outcomes were unobserved and are missing by design. This IRT scaling, in which the item parameters are estimated simultaneously and separate ability distributions are assumed in the two populations while the parameters of the anchor are assumed to be identical in both populations, is usually referred to as "concurrent calibration." Another calibration method is the "fixed parameters method." This approach differs from concurrent calibration in that common items whose parameters are known (either from a previous year calibration or a separate calibration) are anchored or fixed to their known estimates during calibration of other forms. By treating these common item parameters as known, they are not estimated and the item parameters from the uncommon items are forced onto the same scale as the fixed items. This procedure is even more restrictive than concurrent calibration.

For more details the reader is referred to Kolen and Brennan (1995), Stocking and Lord (1983), Haebara (1980), or Hambleton et al. (1991).

3. A Lagrangean Approach to IRT Linking

Let the sample size of the group from P that takes (X, V) be denoted by N_P, let the sample size of the group from Q that takes (Y, V) be N_Q, and denote $N = N_P + N_Q$.

We will use the following notation for the item parameters in the different test forms and populations:

$$\beta_i = \begin{cases} \beta_{X_{Pj}}, & 1 \leq i \leq J, \\ \beta_{V_{Pl}}, & J+1 \leq i \leq (J+L), \\ \beta_{V_{Ql}}, & (J+L)+1 \leq i \leq (J+2L), \\ \beta_{Y_{Qk}}, & (J+2L)+1 \leq i \leq \text{IP}, \end{cases} \tag{1}$$

where IP$= J + 2L + K$ denotes the total number of items, and J, L, and K are the number of the items in the tests X, V, and Y, respectively. For example, $\beta_{X_{Pj}}$ denotes the (possible vector-valued) item parameter for item j from the set of items from the test X that was taken be the examinees from P. Similarly, $\beta_{V_{Pl}}$ denotes the (possibly vector-valued) item parameter for item l from the set of items from the anchor test V that was taken by the examinees from P, and so forth.

The total number of the item parameters (TNIP) is the dimension of the vector of the items parameters times the number of parameters per item. For example, if all items are modelled via the Rasch model, TNIP $= 1 \times$ IP; for a 2PL model, TNIP $= 2 \times$ IP; and for the 3PL model, TNIP $= 3 \times$ IP. For mixtures of item model types in one test, TNIP is the sum of individual item parameter dimensions (1, 2, or 3 for dichotomous items and 2 or more for partial-credit items) over all items.

3.1. Separate Calibration

When estimating separately, the item and ability distribution parameters for population P are maximized given data $(\boldsymbol{X}, \boldsymbol{V}_P)$, separately from the item and ability distribution

parameters for population Q given data $(\boldsymbol{Y}, \boldsymbol{V}_Q)$.

Technically, this can be accomplished by fitting one IRT model to the combined data without assuming that the common items have the same item parameters in both populations.

As mentioned before, the parameter space is described by all the parameters in the IRT model fitted to the data from both populations, in a marginal maximum likelihood (MML) framework.

Let π_P and π_Q denote the parameters used to model the ability distribution. We may think of them as $\pi_{P,Q} = (\mu_{P,Q}, \sigma_{P,Q})$ in the case where we assume normal ability distributions. In somewhat more flexible models, we may assume that the $\pi_{P,Q}$ is a set of multinomial probabilities over quadrature points approximating arbitrary distribution shapes.

Hence, the parameter space is defined by the parameters

$$\boldsymbol{\eta}^t = (\beta_{X_P}, \beta_{V_P}, \pi_P, \beta_{Y_Q}, \beta_{V_Q}, \pi_Q). \tag{2}$$

Given Assumptions 1 and 2 and the properties of the logit and logarithm functions, we can rewrite the joint log-likelihood function for the IRT model applied to the data from both populations as the sum of the two log-likelihood functions, that is,

$$\begin{aligned} L(\boldsymbol{\eta}; \boldsymbol{X}, \boldsymbol{V}_P, \boldsymbol{Y}, \boldsymbol{V}_Q) &= L(\beta_{X_P}, \beta_{V_P}, \pi_P; \boldsymbol{X}, \boldsymbol{V}_P) \\ &\quad + L(\beta_{Y_Q}, \beta_{V_Q}, \pi_Q; \boldsymbol{Y}, \boldsymbol{V}_Q) \end{aligned} \tag{3}$$

In other words, the two separate models are estimated and the two log-likelihood functions are maximized jointly using MML.

Now, it is easy to conceive *any* linking function as a restriction function on the parameter space and any linking process as a maximization of (3) under the linking restriction. Later we will illustrate in detail how this approach works for each linking method. Next we will

illustrate the concurrent calibration method in some detail, then outline how this approach translates to each of the other IRT linking methods: mean-mean, mean-sigma, Stocking and Lord, Haebara, and fixed parameters.

3.2. Lagrangean Concurrent Calibration

When estimating concurrently, the item and ability distribution parameters for population P are maximized given data ($\boldsymbol{X}$, $\boldsymbol{V}_P$) simultaneously with the item and ability distribution parameters for population Q given data ($\boldsymbol{Y}$, $\boldsymbol{V}_Q$).

Technically, two separate ability distributions are estimated, and the two log-likelihood functions are maximized jointly with certain restrictions on the item parameters, namely

$$\beta_{V_{Pl}} = \beta_{V_{Ql}} \tag{4}$$

for $l = 1, \ldots, L$.

Let R denote the L-dimensional restriction function with the components given by

$$R_l(\boldsymbol{\eta}) = k_l(\beta_{V_{Pl}} - \beta_{V_{Ql}}), \tag{5}$$

with $k_l = 1$ for active restrictions on item l.

For the 2PL and 3PL, the restrictions may be imposed only on the b parameters and not on the slope and guessing parameters. This can be achieved by first using projections, h, and then imposing the same constraints, that is, use $b_{Vl} = h(\beta_{Vl})$ and then use $R_l(\boldsymbol{\eta}) = k_l(h(\beta_{V_{Pl}}) - h(\beta_{V_{Ql}}))$.

Hence, the concurrent calibration refers to maximizing (3) under the restriction

$$R_l(\boldsymbol{\eta}) = 0. \tag{6}$$

This setup, maximizing (3) under the restriction (6), is used whenever certain item parameters are assumed to be equal across populations, in our case, across P and Q.

Given a vector $\boldsymbol{\lambda}$ of Lagrangean multipliers, the linking process can be viewed as

maximizing the modified log-likelihood function

$$\begin{aligned}\Lambda(\boldsymbol{\eta}, \boldsymbol{\lambda}; \boldsymbol{X}, \boldsymbol{V}_P, \boldsymbol{Y}, \boldsymbol{V}_Q) &= L(\beta_{X_P}, \beta_{V_P}, \pi_P; \boldsymbol{X}, \boldsymbol{V}_P) \\ &\quad + L(\beta_{Y_Q}, \beta_{V_Q}, \pi_Q; \boldsymbol{Y}, \boldsymbol{V}_Q) - \boldsymbol{\lambda}^t R(\boldsymbol{\eta}). \qquad (7)\end{aligned}$$

Note that if we choose $k_l = 0$ for all l the restriction functions, $R_l(\boldsymbol{\eta})$ are constants, so that instead of concurrent calibration with equality constraints, maximizing the likelihood simultaneously yields separate calibrations and allows the item parameters in the anchor test V to differ between P and Q.

The function in (7) is then maximized with respect to parameters $\boldsymbol{\eta}$ and $\boldsymbol{\lambda}$.

In concurrent calibration, the above equation includes a term R_l for each item $l = 1, \ldots, L$ in the anchor test V. This term enables the imposition of equality constraints on the parameters $\beta_{V\bullet}$.

3.3. Lagrangean Fixed Parameters Scale Linkage

In this method, common items whose parameters are known (for example, from a previous administration calibration or a separate calibration) are anchored or fixed to their known estimates, $w_l,\ l = 1, .., L$, during calibration of other forms. These common item parameters are treated as known and, therefore, they are not estimated; the item parameters from the items that are not common to the forms are forced onto the same scale as the fixed items. This calibration procedure is even more restrictive than concurrent calibration.

As in Section 3.2, let now R denote the $2L$-dimensional restriction function with the components given by

$$R(\boldsymbol{\eta})^t = (R_{Pl}(\boldsymbol{\eta}),\ R_{Ql}(\boldsymbol{\eta})) \qquad (8)$$

where

$$\begin{aligned} R_{Pl}(\boldsymbol{\eta}) &= k_l(\beta_{V_{Pl}} - w_l), \\ R_{Ql}(\boldsymbol{\eta}) &= k_l(\beta_{V_{Ql}} - w_l), \end{aligned} \tag{9}$$

with $k_l = 1$ for active restrictions on item l.

Hence, the concurrent calibration refers to maximizing (3) under the restriction

$$R(\boldsymbol{\eta}) = \mathbf{0}. \tag{10}$$

3.4. Lagrangean Mean-mean IRT Scale Linking

If an IRT model fits the data then, any linear transformation[1] (with slope A and intercept B) of the θ-scale also fits these data, provided that the item parameters are also transformed (see, for example, Kolen & Brennan, 1995, pp. 162–167).

In the NEAT design, the most straightforward way to transform scales when the parameters were estimated separately is to use the means and standard deviations of the item parameter estimates of the common items for computing the slope and the intercept of the linear transformation. Loyd and Hoover (1980) described the mean-mean method, where the mean of the a-parameter estimates for the common items is used to estimate the slope of the linear transformation. The mean of the b-parameter estimates of the common items is then used to estimate the intercept of the linear transformation (see Kolen & Brennan, p. 168).

Lagrange multipliers may also be used to achieve IRT scale linking according to the mean-mean approach. Again, maximizing the modified log-likelihood function Λ given in (7) with a different set of restrictions does the trick. For the mean-mean IRT linking, the restriction function is two-dimensional with the components R_a and R_b, that is, $R^t = (R_a, R_b)$. If we want to match the mean of anchor parameters of population P, we define

$$R_a(\boldsymbol{\eta}) = \left(\sum_{l=1}^{L} h_a(\beta_{V_{Ql}}) - A_P \right) \tag{11}$$

with a constant term $A_P = \sum h_a(\beta_{V_{Pl}})$, which is not viewed as a function of the β_{VP} (but is recomputed at each iteration during maximization) in order to allow unconstrained maximization for P and enforce the mean of β_{VQ} to match this mean in P. As has been explained, h is a projection.

The same is done with the difficulty parameters $b_l = h_b(\beta_l)$, that is,

$$R_b(\boldsymbol{\eta}) = \left(\sum_{l=1}^{L} h_b(\beta_{V_{Ql}}) - B_P \right) \tag{12}$$

This new approach to IRT linking includes also a more general approach that handles populations P and Q symmetrically using

$$R_a(\boldsymbol{\eta}) = \left(\sum_{l=1}^{L} h_a(\beta_{V_{Ql}}) - h_a(\beta_{V_{Pl}}) \right) \tag{13}$$

with $h_a(\beta_i) = a_i$ and

$$R_b(\boldsymbol{\eta}) = \left(\sum_{l=1}^{L} h_b(\beta_{V_{Ql}}) - h_b(\beta_{V_{Pl}}) \right) \tag{14}$$

with $h_b(\beta_i) = b_i$. This avoids the arbitrary choice whether to match P's or Q's slope and difficulty means on the anchor test V.

3.5. Lagrangean Mean-var IRT Scale Linking

The mean-var IRT scale linkage (Marco, 1977) can obviously be implemented in the same way, with only a slight difference in the restrictions used. The means and the standard deviations of the b-parameters are used to estimate the slope and the intercept of the linear transformation.

Again, a two-dimensional restriction function with components R_a and R_b is needed. In order to match the mean and variance of the anchor tests difficulty parameter in population P, we define

$$R_a(\boldsymbol{\eta}) = \left(\sum_{l=1}^{L} h_a(\beta_{V_{Ql}}) - B_P \right) \tag{15}$$

with a constant term $B_P = \sum h_a(\beta_{V_{Pl}})$, which again is not viewed as a function of the β_{V_P}. The same is done with the squared difficulties $b_l^2 = h_b^2(\beta_l)$, that is,

$$R_b(\boldsymbol{\eta}) = \left(\sum_{l=1}^{L} h_b^2(\beta_{V_{Ql}}) - B_P^2 \right) \tag{16}$$

where $B_P^2 = \sum h_b^2(\beta_{V_{Pl}})$.

As before, a more general approach handles populations P and Q symmetrically using

$$R_a(\boldsymbol{\eta}) = \left(\sum_{l=1}^{L} h_b(\beta_{V_{Ql}}) - \sum_{l=1}^{L} h_b(\beta_{V_{Pl}}) \right) \tag{17}$$

with $h_b(\beta_i) = b_i$ and

$$R_b(\boldsymbol{\eta}) = \left(\sum_{l=1}^{L} h_b^2(\beta_{V_{Ql}}) - \sum_{l=1}^{L} h_b^2(\beta_{V_{Pl}}) \right) \tag{18}$$

with $h_b^2(\beta_i) = b_i^2$.

3.6. Lagrangean Stocking and Lord Scale Linkage

Characteristic curves transformation methods were proposed (Haebara, 1980; Stocking & Lord, 1983) in order to avoid some issues related to the mean-mean and mean-var approaches. For the mean-mean and mean-var approaches, various combinations of the item parameter estimates produce almost identical item characteristic curves over the range of ability at which most examinees score.

The Stocking and Lord IRT scale linkage finds parameters for the linear transformation of item parameters in one population (say Q) that matches the test characteristic function of the anchor in the reference population (say P).

The Stocking and Lord transformation finds a linear transformation (i.e., a slope A and an intercept B) of the item parameters—difficulties and slopes—in one population based on a matching of test characteristic curves. Expressing this in the marginal maximum

likelihood framework yields

$$(A, B) = \min \left[\sum_{g=1}^{G} \pi_g^* \left(\sum_{l=1}^{L} p\left(\theta_g, \beta_{VPl}\right) - p\left(\theta_g, B + A\beta_{VQl}\right) \right)^2 \right] \tag{19}$$

where the weights π_g^* of the quadrature points θ_g for $g = 1, \ldots, G$ are given by

$$\pi_g^* = \frac{n_{Pg}\pi_{Pg} + n_{Qg}\pi_{Qg}}{n_{Pg} + n_{Qg}}. \tag{20}$$

We propose using a method employing the same rationale as the Stocking and Lord approach, namely optimizing the match of the test characteristic curves between the anchors V_P and V_Q. In the proposed framework, the primitive of these functions, which is the criterion to be minimized to match the two test characteristic functions as closely as possible, is defined as

$$R^{SL}(\boldsymbol{\eta}) = \left[\sum_{g=1}^{G} \pi_g^* \left(\sum_{l=1}^{L} p\left(\theta_g, \beta_{VPl}\right) - p\left(\theta_g, \beta_{VQl}\right) \right)^2 \right]. \tag{21}$$

In order to minimize (21), we implement the Lagrangeans in such a way that

$$\Lambda(\boldsymbol{\eta}, \lambda) = L(X, V_P) + L(Y, V_Q) - \boldsymbol{\lambda} \mathbf{J}_{R^{SL}}(\boldsymbol{\eta}) \tag{22}$$

or, more explicitly,

$$\Lambda(\boldsymbol{\eta}, \lambda) = L(X, V_P) + L(Y, V_Q) - \lambda_P^T \mathbf{J}_{P, R^{SL}}(\boldsymbol{\eta}) - \lambda_Q^T \mathbf{J}_{Q, R^{SL}}(\boldsymbol{\eta}). \tag{23}$$

The components of interest of the Jacobian $\mathbf{J}_{R^{SL}}(\boldsymbol{\eta})$ are defined by components for the anchor item parameters in P

$$\frac{\partial R^{SL}(\boldsymbol{\eta})}{\partial \beta_{i,P}} = \sum_{g=1}^{G} \pi_g^* \frac{\partial p\left(\theta_g, \beta_{VPi}\right)}{\partial \beta_{iVP}} 2 \left[\sum_{l=1}^{L} p\left(\theta_g, \beta_{VPl}\right) - p\left(\theta_g, \beta_{VQl}\right) \right], \tag{24}$$

and for the components representing the item parameters in Q we find

$$\frac{\partial R^{SL}(\boldsymbol{\eta})}{\partial \beta_{i,Q}} = -\sum_{g=1}^{G} \pi_g^* \frac{\partial p\left(\theta_g, \beta_{VQi}\right)}{\partial \beta_{iVP}} 2 \left[\sum_{l=1}^{L} p\left(\theta_g, \beta_{VPl}\right) - p\left(\theta_g, \beta_{VQl}\right)\right] \tag{25}$$

because of the negative sign of all $\beta_{\bullet,Q}$ terms. The derivatives in the equations above actually represent vector-valued derivatives if $\beta_{i,P|Q}$ is vector-valued. For example, we have

$$\frac{\partial R^{SL}(\boldsymbol{\eta})}{\partial \beta_{i,P|Q}} = \left(\frac{\partial R^{SL}(\boldsymbol{\eta})}{\partial a_{i,P|Q}}, \frac{\partial R^{SL}(\boldsymbol{\eta})}{\partial b_{i,P|Q}}, \frac{\partial R^{SL}(\boldsymbol{\eta})}{\partial c_{i,P|Q}}\right)$$

in the case of the 3PL model.

By maximizing (20), we will find the transformation of the difficulties and the slopes in one population based on matching test characteristics. Note that in our approach this transformation need not to be linear (although it will be linear if the model fits the data).

3.7. Lagrangean Haebera Scale Linkage

Haebara (1980) expressed the differences between the characteristic curves as the sum of the squared differences between the item characteristic functions for each item over the common items for examinees of a particular ability θ_n. The Haebara method is more restrictive than the Stocking and Lord method because the restrictions take place at the item level (i.e., for each item), while the Stocking and Lord approach poses a global restriction at the test level.

The slope and the intercept of the linear transformation can be found by minimizing the expression on the right-hand-side of (26):

$$(A, B) = \min \left[\sum_{g=1}^{G} \pi_g^* \sum_{l=1}^{L} \left(p\left(\theta_g, \beta_{VPl}\right) - p\left(\theta_g, B + A\beta_{VQl}\right)\right)^2\right], \tag{26}$$

(see, for example, Kolen & Brennan, 1995, p. 170).

The algorithm we are proposing is similar to the one described previously for the Stocking and Lord scale linking; the only difference (from the computational point of view)

is in the form of the restriction function, that is:

$$R^H(\boldsymbol{\eta}) = \left[\sum_{g=1}^{G} \pi_g^* \sum_{l=1}^{L} \left(p\left(\theta_g, \beta_{VPl}\right) - p\left(\theta_g, \beta_{VQl}\right)\right)^2\right]. \tag{27}$$

As before, in order to minimize (27), we implement the Lagrangeans in such a way that

$$\Lambda(\boldsymbol{\eta}, \lambda) = L(X, V_P) + L(Y, V_Q) - \boldsymbol{\lambda}\mathbf{J}_{R^H}(\boldsymbol{\eta}) \tag{28}$$

or, more explicitly,

$$\Lambda(\boldsymbol{\eta}, \lambda) = L(X, V_P) + L(Y, V_Q) - \lambda_P^T \mathbf{J}_{PR^H}(\boldsymbol{\eta}) - \lambda_Q^T \mathbf{J}_{QR^H}(\boldsymbol{\eta}) \tag{29}$$

where the components of interest of the Jacobian $\mathbf{J}_{R^H}(\boldsymbol{\eta})$ are

$$\frac{\partial R^H(\boldsymbol{\eta})}{\partial \beta_{lP}}, \frac{\partial R^H(\boldsymbol{\eta})}{\partial \beta_{lQ}}. \tag{30}$$

Again, the partial derivatives may be vector-valued for each $\beta_{i,P|Q}$, so that the dimension of the restriction is approximately $2L$ times the average number of item parameter dimensions, 3 if only the 3PL model is used but possibly higher when generalized partial-credit items or other polytomous item response models are present.

4. Discussion

This paper presents a new perspective on IRT linking. It introduces a unified approach to IRT linking, emphasizing the similarities between different methods. We show that IRT linking might consist of a family of IRT linking functions, where restrictions can be "turned on or off," according to what the data might suggest. Moreover, this new approach allows both generalizations and exactly matching implementations of the existing methods, since the existing IRT linking methods are included as special cases in this new family of IRT linking functions.

We believe that this approach will allow the development of statistical tests (such as

Lagrange multiplier tests) for checking the appropriateness of different IRT linking methods (see Glas, 1999, for a similar approach used for investigating nested IRT models). Such a test would allow to check whether lifting certain restrictions will yield a significant improvement in model-data fit, for example in a case where Lagrangean concurrent calibration is used for all anchor items in a vertical linkage and a certain set of items exposes parameter drift over time.

This approach to IRT linking can be easily viewed in an MCMC framework, where, by specifying appropriate prior distributions, the estimation of the modified likelihood functions is straightforward.

At the same time, the view of any linking function as a restriction function implies a larger flexibility in the linking process: when dealing with vertical linking, this method can incorporate the modelling of growth, possibly expressed as a hierarchical structure of the item parameters in the anchor.

Such a hierarchical structure was proposed by Patz et al.(2003); they used the MCMC estimation method. The hierarchical approach proposed by Patz et al. is "a more general version of concurrent estimation of the unidimensional IRT model" (p. 40), and their motivation has similarities with ours: to unify the two most commonly used linking methods for vertical equating, the very restrictive concurrent calibration method and separate calibration followed by a test characteristic curves linking.

If we recast this hierarchical approach of the proficiencies across grades into a hierarchical structure of the (common) item difficulties, a short summary of the Patz et al. (2003) approach (slightly generalized) in our notation is

$$R_l(\boldsymbol{\eta}) = k_l(h(\beta_{V_{Pl}}) - f(h(\beta_{Q_{Pl}}))), \tag{31}$$

where $k_l = 1$ for active restrictions on item l, R_l denotes the component l of the restriction function, h is the projection described before, and f is a function of the common item parameters of the old administration (or previous grade). In order to obtain the hierarchical

structure at the level of the difficulties of the common items, we consider

$$h(\beta_{V_{Pl}}) = b_{V_{Pl}}.$$

Following the approach of Patz et al. (2003), the relationship between the difficulty of the item parameters across grades can be expressed as a quadratic function,

$$f(h(\beta_{Q_{Pl}})) = f(b_{Q_{Pl}}) = \alpha b_{Q_{Pl}}^2 + \gamma b_{Q_{Pl}} + \delta, \tag{32}$$

where α, γ, δ are additional parameters of the model that need to be estimated. Furthermore, the modified likelihood function, with a restriction function described in (31) can be maximized using the Lagrange multipliers in the same way as explained for the other linking methods.

Note that from a computational point of view, this is only a slight generalization of the restriction functions described for the mean-mean and mean-var linking methods.

Obviously, additional investigations are necessary in order to insure that the model is identified and to insure the convergence of the maximization algorithm. Although, here we propose an analytical approach and we will try to use an expectation-maximization (EM) algorithm, an MCMC estimation method would be straightforward to implement.

Moreover, the approach presented in this paper may easily be extended to multidimensional IRT models, at least for simple structure multiscale IRT models (like the one used in NAEP and other large scale assessments); there is no additional formal work necessary and the method proposed in this report can be readily applied. Patz et al. (2003) also investigated multidimensional IRT models for vertical linking and used the MCMC estimation method. However, implementing and maximizing modified likelihood functions under such restrictions using analytical methods are of interest for future research.

Longitudinal studies might also benefit from these two approaches: one that assumes a hierarchical structure in the item parameters of the anchor and/or one that assumes a multidimensionality of proficiencies (or of common item difficulties) across school grades.

This flexibility may also be a desirable feature in educational large-scale assessments, where in some instances it is necessary to relax the restriction of equality of all item parameters. In conclusion, this new approach is very promising for assessment programs that use IRT linking.

References

Aitchison, J., & Silvey, S.D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics, 29*, 813–828.

Angoff, W.H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: ETS. (Reprinted from *Educational measurement,* 2nd ed., pp. 508–600, by In R.L. Thorndike (Ed.), 1971, Washington, DC: American Council on Education.)

von Davier, A.A. (2003). *Large sample tests for comparing regression coefficients in models with normally distributed variables* (ETS RR-03-19). Princeton, NJ: ETS.

von Davier, A.A., Holland, P.W., & Thayer, D.T (2004). *The kernel method of test equating*. New York: Springer Verlag.

Glas, C.A.W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika, 64*(3), 273–294.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research,22,* 144–149.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Kolen, M.J., & Brennan, R.J. (1995). *Test equating: methods and practices*. New York: Springer-Verlag.

Lord, F.M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

Loyd, B.H., & Hoover, H.D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179–193.

Marco, G.L. (1977). Item characteristic curves solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139–160.

Marco, G.L., Petersen, N.S., & Stewart, E.E. (1983). A test of the adequacy of curvilinear score equating models. In Weiss, D.J. (Ed.), *New horizons in testing* (pp. 147–176). New York: Academic Press.

Patz, R., Yao, L., Chia, M., Lewis, D., & Hoskens, M. (2003). *Hierarchical and multidimensional models for vertical scaling*. Paper presented at NCME 2003, April, Chicago, IL.

Petersen, N.S., Marco, G.L., & Stewart, E.E. (1982). A test of the adequacy of linear score equating models. In P.W. Holland & D.B. Rubin (Eds.), *Test equating* (pp. 71–135). New York: Academic Press.

Stocking, M., & Lord, F.M.(1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210.

Thissen, D., & Wainer, H. (2001). *Test scoring*. Mahwah, NJ: Erlbaum.

Notes

[1]A more general result holds: All strictly monotone transformations of θ are also permissible. This feature, however, will not be pursued further in the current paper.